A NEW APPROACH TO URBAN ROAD SAFETY INVOLVING PEDESTRIAN SAFETY ESTIMATION

A. P. AURICH & R. MAIER
Technical University of Dresden, Faculty of Transport and Traffic Sciences, Chair of Road Traffic Engineering and Theory of Transportation Planning, Germany
ALLAN.AURICH@TU-DRESDEN.DE, INFO@QSV-DRESDEN.DE

ABSTRACT

Attempts to predict accidents involving pedestrians in road networks in most cases are faced with the problem of missing pedestrian counts. A comparison of two different procedures is carried out in order to overcome this problem. Missing pedestrian traffic counts are substituted with environmental variables derived from the adjacent building structure and the surrounding land-use within an area of influence. While the building structure is raised alongside each road segment socio-economic and land-use data are derived within a range of 300 m of each segment. A procedure is developed in order to obtain and aggregate the relevant spatial variables by matching geographical databases. Subsequently the correlation structure of the spatial data is analysed. A principal component analysis (PCA) is carried out revealing underlying components and reducing the dimensionality of the initial dataset. Five generalized linear models (GLMs) are computed each with Poisson distributed and negative binomial error structure in order to quantify the interrelations between environmental variables and pedestrian accidents. Models based on the adjacent building structure and those based on principal components of the surrounding land-use reach a similar goodness-of-fit. A model combining building structure and the sale areas within reach of 300 m proves to be the best fit model.

1. INTRODUCTION

Accident figures are the favoured criteria to quantify road safety. Thorough traffic planning implies the assessment of the performance as well as forecasting the safety effects of constructive and conceptual measures in road networks. Accident Prediction Models (APMs) are widely used to handle the latter task. Though safety research has elicited a multitude of models, the best part of these, predict solely collision counts of motorized vehicles.

At the same time non-motorized road-users are involved in more than half of the personal injury accidents in German cities. As an example, approximately two-thirds of all people severely injured or killed in road accidents in the city of Dresden in the year 2005 were cyclists or pedestrians. Hence the consideration of accidents involving non-motorized road-users is crucial for a valid road safety assessment in cities.

Most existing APMs, therefore, seem hardly suitable to determine sufficiently an average safety degree for urban roads. One main reason for the difficulty of predicting non-motorized accidents on road segments is the lack of exposure data. While traffic counts of motorized vehicles are available for arterial networks in most cities, pedestrian and bicycle counts are not. Non-motorized traffic counts are time-consuming and costly since they are difficult to automate. As a result, these data exist for

particular cross-sections at best and do not fulfil the requirements of a sufficient basis for statistical models of road networks. Using motorized traffic counts alone as exposure for accident prediction falls short insofar as non-motorized accident involvement strongly depends on the appearance of these modes.

This paper focuses on accidents involving pedestrians. It deals with the possibility of partly substituting missing information about pedestrian counts with land-use adjacent to road segments. This method is predicated on the basic principle of a strong relationship between land-use and traffic generation.

2. PREVIOUS WORK

Land-use and its socio-economic factors form the basis of all traffic generation models. Thus, the relationship between settlement structure and traffic generation has been thoroughly analysed over the past years. Estimates of generated traffic for German conditions can be obtained from the pertinent manual by the German Road and Transportation Research Association (FGSV) [1]. According to the manual, the basic criteria of type and intensity of land-use are population, housing, social infrastructure (e.g., schools), work-related infrastructure (e.g., retail jobs) and special uses (e.g., cinemas). The traffic data derived from the procedure described in the manual can, in fact, only be regarded as rough estimates since the values differ over a wide range. For example, the expected number of customers per square meter of sales area in shopping malls varies between 30 and 150. In addition, no exact information is given about the expected choice of mode according to the type of land-use.

An & Chen [2] developed an estimation procedure for non-motorized travel demand based on a multivariate regression analysis of infrastructural and census data on a block level. The correlations between a number of socio-economic, environmental, and infrastructural factors and the number of the non-motorized share of the daily commute were analyzed for the city of Lexington, Kentucky. The ensuing regression model shows the strongest predictive power of employment density, percentage of student population, median household income and average sidewalk length together. A more general finding of the analysis was that several socio-economic parameters are highly correlated and therefore have to be treated carefully when used in regression models in order to avoid multicollinearity.

Alrutz & Bohle [3] dealt with pedestrian counts on a finer spatial level and therefore analysed the interrelation of pedestrian counts and the use of adjacent buildings on urban road segments in several German cities. They classified the surroundings with the help of an “urban density” (ratio of built-up street length and total street length along the segment multiplied by the number of storages) as well as the “structure of use” (ratio of built-up length including retail and the entire built-up length along the segment). The correlation results showed large effects between pedestrian traffic counts and both the “structure of use” ($r = 0.79$) and “urban density” ($r = 0.56$).

Their results were taken up by Monse [4], who investigated several parameters in relation to traffic counts and accident counts with pedestrians as well as cyclists involved in the arterial urban road network of the city of Dresden (54 segments). In order to describe the urban use of each segment the author combined the “structure of use” and the “urban density” by multiplying their values. The analysis showed sig-

nificantly high correlations between pedestrian accidents and urban use (r = 0.70) while the relationship between urban use and bicycle accidents was significantly weaker (r = 0.33). The average bicycle travel distances exceed those of pedestrians. From this point of view it seems logical that the land-use adjacent to a road has a lesser influence on bicycle accidents than accidents involving pedestrians.

A first attempt to include a parameter similar to those introduced by Alrutz & Bohle [3] and Monse [4] in a generalized regression framework was carried out by Schueller [5], who determined the influence of free speeds on the frequency and severity of accidents on urban major roads in Dresden. Schueller once again modified the approach by differentiating the built-up length by four different structures (residential, retail, combined residential and retail, industrial). A parameter (KLF) was computed for each segment between two major Intersections in the following manner:

$$KLF = \frac{L_{ind} + 2 \cdot L_{res} + 2 \cdot L_{ret} + 3 \cdot L_{res,ret}}{2 \cdot L_{seg}}$$

where:
KLF = parameter characterizing adjacent structure of building uses
L_{ind} = built up length with industrial use (total of both sides)
L_{res} = built up length with residential use (")
L_{ret} = built up length with retail (")
$L_{res,ret}$ = built up length combining residential use and retail (" ; in the same building)
L_{seg} = segment length

The weighting factors used in the formula were obtained through preliminary regression analyses for a data set of road segments in Dresden. Schueller used the KLF values for predicting accidents with personal injury as well as nonmotorized accidents on a reduced dataset of road segments. A more specific use for analysing traffic safety of pedestrians within the entire main road network is yet to follow.

3. OBJECTIVE AND METHODOLOGY

The objective of this analysis is to show and compare possible approaches of incorporating land-use data in safety models and surveys in order to improve the predictive power for non-motorized accidents. Besides the above-mentioned procedure of classifying the length and use of buildings situated alongside the road segment, an alternative approach is developed using socio-economic and land-use data of the area surrounding each segment.

The benefit of this second approach lies in the availability of data. Experience with German data shows that information about the exact use of buildings is rarely available, especially if retail and residential uses are combined in one building. For this reason the use of the built-up lengths in most cases has to be assessed through analysing aerial images or on-site surveys. On the other hand, land-use and socio-economic data are normally collected as a basis for traffic generation models. The use of the same database ensures a high applicability and enables further development as well as automation in the scope of future traffic planning processes.

Two different types of environmental data are therefore compared:

- Information concerning buildings adjacent to road segments and their use.
- Socio-economic and land-use data within a defined distance of road segments.

In a first step, the geographical data from the different sources are processed in order to obtain a network model merging infrastructural, traffic-related, socio-economic, land-use and accident data. Subsequently the focus will lie on the methodology of assigning the socio-economic and land-use data (spatial) to the respective road segments (linear).

The spatial indicators derived are analysed in respect of their interdependencies by computing correlations and running a Principal Component Analysis (PCA). Finally, the parameters characterizing adjacent urban structure are included in Generalized Linear Models (GLMs) in order to quantify the interrelations with non-motorized accident counts.

4. DATA

4.1. Road network

The study is deals with the urban main road network of Dresden, a German city with a population of approximately 500,000.

In a first step, the arterial road network is subdivided into road segments and intersections. Intersections of two or more main roads are not the subject of this survey and, therefore, are excluded from the digital network. Preliminary analyses showed a considerable influence of intersections on the number of accidents over an adjacent road length of 50 m. In order to avoid bias in the models the first 50 m of road length next to intersections were also omitted from the network.

Also accidents at intersections with minor roads are excluded from the sample. Further analyses show only marginal effects of these minor intersections beyond their geographical boundaries. For this reason no further adjacent segment length was omitted as at major intersections. If infrastructural variables do not change before and after minor intersections the segment is not subdivided.

The variables concerning traffic and road infrastructure recorded for each segment are listed in Table 1. In cases where any of these variables change between two major intersections, the initial segment is divided into further sub-segments.

Table 1 – Traffic and Road infrastructure variables for road segments

variable	description	level of measurement
length	segment length (km)	continous
AADT	average annual daily traffic (veh./24h)	continous
Tram	tram on segment (yes / no)	dichotomous
TramNo	number of trams per day (veh./24h)	continous
TramStops	daily number of trams stopping at stations (100 stops/24h)	continous
BusStops	daily number busses stopping at Bus-stops (100 stops/24h)	continous
lanes	numbers of lanes (2 or less / 3 or more)	dichotomous
median	central reserve (yes / no)	dichotomous
BikeFac	bicycle facility (bicycle path / bicycle lane / none)	categorial
Park	parking on the segment (yes / no)	dichotomous

By nature, accidents are statistically rare occasions and counts only appear as integers. Consequently, short segments lead to large variations of accidents per kilometre. Segments shorter than 50 m were therefore also excluded from the database.

As a result, 668 segments with a total road length of 283 kilometres remain for the subsequent analyses. Trams travel on roughly 87 kilometres of this total road length.

4.2. Accidents

Accident counts are taken from the digital database maintained by the police. The analysis includes all accidents with pedestrians or with bicycle involvement during the time period 2004 to 2008.

A total of 536 accidents involving pedestrians (470 personal injury accidents) occurred on the remaining road segments during the five-year-period.

4.3. Socio-economic and land-use data

Two different geographical databases are used for generating the socio-economic and land-use indicators: On the one hand, the digital geographical model (DLM) provided by the land surveying office of Saxony; on the other, the digital geographical basis of the traffic generation model of the municipality of Dresden. The DLM contains information concerning boundaries of plots of land and in some cases information about their land-use (e.g., schools, universities, residential use, and railway stations).

The database of the traffic model contains the boundaries of 529 traffic analysis zones (TAZs) and the corresponding socio-economic and land-use data. The data drawn from the TAZ database include population numbers, numbers of workplaces, school and university places as well as sales areas. The population numbers are disaggregated according to patterns of traffic-related behaviour (VHG). Table 2 contains a detailed list of variables extracted in the following steps.

Table 2 – Data drawn from the traffic analysis zones

attribute	description
VHG 1	population number age 0-5 years
VHG 2	population number age 5-17 years
VHG 3	population number age 18-64 years, unemployed, no car available
VHG 4	population number age 18-64 years, unemployed, car available
VHG 5	population number age 18-64 years, employed, no car available
VHG 6	population number age 18-64 years, employed, car available
VHG 7	population number age 65+ years, no car available
VHG 8	population number age 65+ years, car available
Pop	total population number
Emp	number of employed
Stud	number of students
WoSer	number of workplaces - service
WoPro	number of workplaces - productive
Scho	school places
Kind	Kindergarten places
Sale	sales area (m^2)
Cars	number of registered cars

First of all, both databases are matched geographically by allotting each built-up plot of land to the corresponding traffic analysis zone. The socio-economic and land-use data are then proportionally redistributed over all plots within each TAZ. The number of school and university places are assigned to plots with these specific uses, while all other data are distributed over the remaining plots (mixed use). By utilizing the aggregated plots, instead of the entire area of the TAZ, a more plausible distribution of data is achieved. This way unexploited areas and those with unsuitable land-use are excluded from the determination. As an additional consequence, the bias caused by inaccurate geographical information is reduced.

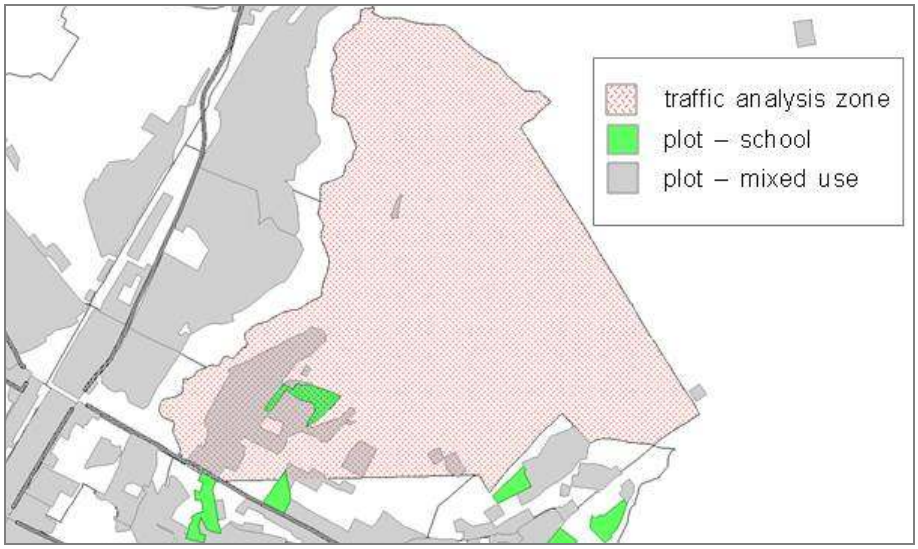

Figure 1 – Matching of traffic analysis zones and built-up plots of land

The principle is shown in Figure 1 using the example of a selected traffic analysis zone (marked red). Without projecting the data onto the related plots of built-up land, the data would be uniformly spread across the area of the TAZ. Subsequent steps of

analysis concerning data within a certain range of corresponding road segments would, thus, be most certainly biased.

The allocation of the zonal attributes to corresponding road segments follows the basic idea of analysing the potential of non-motorized traffic within a sphere of influence. The attributes are extracted by generating concentric buffers around each road segment with an appropriate radius. A convenient distance is therefore chosen at 300 m, referring to a common accessibility criterion for bus-stops.

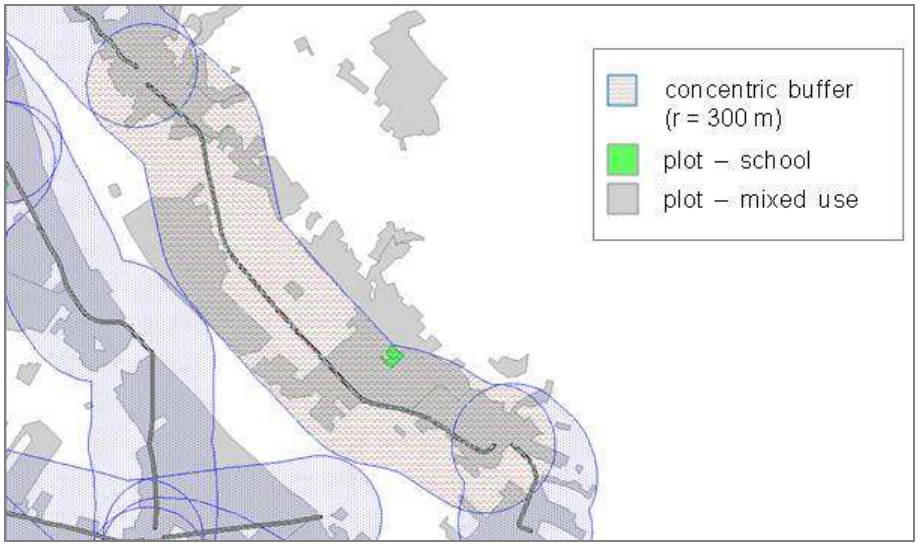

Figure 2 – Allocation of zonal attributes with concentric buffers

The zonal attributes are computed proportionally according to the ratio of the part of the plot within the buffer and the entire area of the plot:

$$X_{buf} = \sum_{i=1}^{n} \left(\frac{A_{i,buf}}{A_i} \cdot X_i \right)$$

where:
X_{buf} = total attribute allocated to buffer
$A_{i,buf}$ = fraction of plot “i" within buffer
A_i = total area of plot “i"
X_i = total attribute of plot “i"

The attributes allocated are then transformed to densities through division by the segment length in order to account for varying road lengths. However, this method bears a problem in that short segments will have proportionally higher densities, as every segment buffer draws a semicircle at each end of the segment. The area of this part of the buffer is constant and therefore has a greater effect on short segments than on long ones. This aspect will have to be kept in mind when interpreting the model results.

The use of buffers leads to multiple allocations of data since the buffers overlap. In common four-step traffic demand models, each trip is uniquely assigned to a certain route, i.e., road link. Within the approach described this is not the case, due to these overlaps. The socio-economic and land-use data derived within the sphere of influence can therefore not be interpreted as in the scope of traffic generation, but rather

as a non-motorized potential. This inaccuracy is due to the need for economy when carrying out a network-wide analysis and is knowingly accepted. As a consequence, the variance of the related accident counts is most likely to increase whereas the diversity of the allocated socio-economic and land-use data among the road segments decreases. This, again, has to be considered when interpreting the results of the PCA as well as the GLMs.

5. STATISTICAL DATA ANALYSIS

5.1. Correlation analysis

The Pearson correlations are computed between the individual socio-economic and land-use density variables within the sphere of influence (300 m) of road segments. In addition the variable KLF is included in the analysis. The corresponding correlation matrix is shown in Table 3.

Table 3 – Correlation matrix of socio-economic and land-use variables (densities)

Pearson Correlation	VHG 1	VHG 2	VHG 3	VHG 4	VHG 5	VHG 6	VHG 7	VHG 8	Emp	Pop	WoSer	WoPro	Kind	Scho	Sale	Cars	KLF
VHG 1	1																
VHG 2	,936**	1															
VHG 3	,767**	,679**	1														
VHG 4	,806**	,767**	,965**	1													
VHG 5	,906**	,860**	,901**	,883**	1												
VHG 6	,908**	,955**	,728**	,833**	,878**	1											
VHG 7	,367**	,509**	,467**	,508**	,569**	,567**	1										
VHG 8	,385**	,570**	,385**	,486**	,512**	,638**	,956**	1									
Emp	,933**	,950**	,805**	,873**	,942**	,988**	,584**	,614**	1								
Pop	,873**	,907**	,853**	,911**	,938**	,951**	,719**	,724**	,973**	1							
WoSer	,103**	,047	,388**	,306**	,283**	,077*	,405**	,252**	,147**	,261**	1						
WoPro	,075	-,014	,118**	,109**	,101**	,066	,064	,020	,080*	,081*	,374**	1					
Kind	,505**	,553**	,455**	,508**	,520**	,563**	,395**	,404**	,564**	,567**	,133**	,025	1				
Scho	,111**	,109**	,392**	,382**	,226**	,147**	,260**	,187**	,177**	,264**	,217**	,005	,197**	1			
Sale	-,037	-,019	,048	,043	,018	-,004	,160**	,130**	,003	,049	,487**	-,004	,032	-,042	1		
Cars	,819**	,913**	,666**	,763**	,839**	,943**	,729**	,777**	,935**	,940**	,220**	,084*	,556**	,168**	,056	1	
KLF	,557**	,537**	,475**	,497**	,530**	,528**	,281**	,291**	,543**	,532**	,159**	,041	,378**	,131**	,049	,508**	1

* p < 0.05, ** p < 0.01

The highest correlations can be noted between the individual population groups (VHG 1–8, Pop, Emp) as well as the number of registered cars (Cars), Kindergarten places (Kind) and all population variables. These high correlations are expectable and have to be kept in mind for further regression analyses. In order to avoid multicollinearity either only one of these variables should be included at a time, or the model design will have to account for interactions.

Both the workplace-related variables correlate significantly with population variables, whereby the densities of service-related workplaces (WoSer) show higher values than the production-related (WoPro). Though service-related workplaces mostly correlate significantly, the coefficients do not reach very high values. It should be considered, that even small correlations become significant at a sample size of

n = 668. The highest coefficient results from the correlation between the densities of service-related workplaces and sales area (Sale), at approximately 0.5. Densities of production-related workplaces seem to be distributed independently from further variables considered.

The densities of sales area (Sale) only show a few noteworthy correlations. Except for its correlation with service-related workplaces (WoSer), this variable can also be regarded as distributed independently from the othervariables analysed.

The structure of the adjacent buildings (KLF) correlates significantly with all variables except for production-related workplaces (WoPro) as well as sale area (Sale). A combined regression model including the zonal attributes as well as the structure of adjacent buildings should therefore be limited to these three variables besides variables concerning road infrastructure and traffic parameters.

5.2. Principal component analysis (PCA)

The aim of PCA is to find a set of latent variables underlying the data observed. It can be used for a reduction of dimensionality in that an original set of variables is transformed into a substantially smaller set of underlying factors, so called principal components. Depending on the chosen type of transformation the components can be constructed orthogonally. Hence the use of such principal components instead of the original variables prevents multicollinearity in a subsequent regression model. On the downside components are not self-explanatory which may lead to uncertain explanations in some cases. Moreover every kind of data reduction leads to a loss of information. For detailed information the reader is referred to Jolliffe [6].

In the case at hand the procedure is based on the Pearson correlations of the original variables. To allow a valid PCA, the correlation structure has to meet certain requirements. In the present case, the suitability of the data is determined with the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO), Haitkovsky's X^2, and Bartlett's test.

For conceptual reasons and in order to fulfil the requirements the number of variables analysed is reduced to five socio economic and land-use density variables: population (Pop), employment (Emp), registered cars (Cars), total workplaces (Wor, sum of WoSer and WoPro), and sales area (Sale). With this combination both Bartlett's test and Haitkovsky's X^2 are highly significant ($p < 0.001$) while the KMO attains a value of 0.7 ("mediocre" according to [7]).

The underlying components are computed as linear composites of the original variables obtained through orthogonal transformation. The transformation is defined in such way that the variance of the first principal component is maximized. Thus it accounts for the most possible variability within the original data. Each following component is then constructed in the same manner under the constraint of being

orthogonal to the former components. After defining the components, a subsequent rotation can be taken out in order to enable a differentiated interpretation of the components. An orthogonal rotation (VARIMAX) was chosen in this case. The suitable number of components is estimated with the help of so called scree plots and eigenvalue criteria.

The original set of five variables is reduced to two components (C1, C2). Together both components account for 87 % of the variation within the initial variables. The first component (C1) is highly related to the population variables (Pop, Emp, Cars) and can therefore be interpreted as “population component”. The second component (C2) is mainly affected by the economic variables (Wor, Sale), being referred to as “economic component” within the further analysis.

It is important to state that the results of a principal component analysis cannot be extrapolated beyond the analysed sample. Generalization can only be achieved by revealing the same component structure in different samples.

6. GENERALIZED LINEAR MODELS

In general accidents are rare random events that are assumed to be Poisson distributed. Hence regression models based on normally distributed errors and assuming homoscedasticity are inappropriate for analysing accident frequency. In safety research it has therefore become common practice to use generalized linear models (GLM).

Generalized linear models overcome the restrictions of the general linear model in that the stochastic component (error term) can follow other distributions than the normal and the link between the stochastic and the systematic component can be a function other than identity. The error distribution can be any member of the exponential family. The reader is referred to McCullagh & Nelder [7] for detailed information about generalized linear models.

It appears that pure Poisson models of accident counts tend to be affected by overdispersion. In these cases the variance exceeds the expected value. According to Maher & Summersgill [8] there are several possible reasons for this phenomenon: unobserved, explanatory, variables effectively adding to the random error, errors in the explanatory variables, and mis-specified models. Regardless of the possible reasons, overdispersion leads to erroneous parameter estimates as well as confidence intervals and should therefore be accounted for.

It has therefore become state-of-the-art to use models based on a negative binomial error structure (NB models, also called Poisson-gamma models). The negative binomial distribution can be regarded as a combination of a Poisson and a gamma distribution. In this case the Poisson distribution accounts for the variations due to the ran-

dom nature of accidents, while the variations caused by unobserved variables are expected to be gamma distributed.

A logarithmic link function is chosen for the models. The link function can be derived from the exponential form of the Poisson distribution. The resulting log-linear structure ensures non-negative values and therefore best accounts for the characteristics of count data. This leads to the following general model equation:

$$\ln(\mu) = \eta = \alpha + \beta_i X_i \quad \text{or else} \quad \mu = e^{(\eta)} = e^{(\alpha+\beta_i X_i)}$$

where:
μ = expected number of accidents (per segment and year)
η = linear predictor
α = intercept
β_i = coefficients
X_i = variables

5.3. Modelling approach

All models relate to the total number of pedestrian accidents on each segment within the five-year-period. In the first step a model is developed including the traffic-related and infrastructural variables listed in Table 1. Starting from this point, subsequently three models are developed differing in the type of environmental data used to substitute missing pedestrian counts: the structure of adjacent buildings along the road segments (KLF), land-use and socioeconomic data according to Table 2 and the principal components derived from land-use and socioeconomic data within the sphere of influence of each segment. Finally the best model is generated by combining different types of environmental variables with respect to the former correlation analysis.

Both Poisson and negative binomial models are computed. Inspection of the Poisson model allows an assessment of the unexplained variance due to unobserved explanatory variables. On the other hand the coefficients and confidence intervals computed with the NB-models are regarded as being more reliable.

The variables are inserted stepwise, beginning with the null-model only containing the intercept. The decision of whether a variable is to be retained or omitted from the model is made at a significance level of 95 %. Exposure data (length, AADT) are included as logarithms in order to ensure the condition of zero accidents without traffic and at a segment length of zero. Even though a linear relationship between number of accidents and segment length seems logical, preliminary analyses have shown a decreasing accident density with increasing length.

The models containing the different types of variables are compared with the help of Akaike's information criterion (AIC) as well as Pearson's X^2 statistic. Akaike's infor-

mation criterion is based on the maximized likelihood. The AIC not only rewards goodness-of-fit but also penalises the number of estimates included in the model. A detailed explanation is given in Burnham & Anderson [10]. In the case of an adequate model the ratio of Pearson's X^2 statistic and the degrees of freedom is supposed to be close to one. Values exceeding one indicate overdispersion.

7. RESULTS

The model results are listed in Table 4.

Table 4 – Accident models of accidents involving pedestrians

model (U_{pe} =538, n=668)	parameter	coefficient[a]		standard error	95% - confidence interval min.	95% - confidence interval max.	p[b]	negative binomial distribution parameter	negative binomial distribution Pearson X^2	negative binomial distribution AIC	Poisson distribution Pearson X^2	Poisson distribution df	Poisson distribution X^2/df
null	intercept	-1.830	***	0.075	-1.977	-1.682	-	2.525	667.16	1621	2018.85	667	3.027
basic model	intercept	-7.574	***	0.920	-9.378	-5.770	< 0.001	0.800	663.77	1421	1090.15	664	1.642
	ln(length)	0.543	***	0.082	0.383	0.704	< 0.001						
	ln(AADT)	0.626	***	0.097	0.436	0.817	< 0.001						
	TramStops	0.229	***	0.026	0.179	0.280	< 0.001						
adjacent buildings	intercept	-8.642	***	0.918	-10.441	-6.842	< 0.001	0.590	662.90	1379	956.04	663	1.442
	ln(length)	0.668	***	0.082	0.506	0.829	< 0.001						
	ln(AADT)	0.673	***	0.095	0.486	0.860	< 0.001						
	TramStops	0.178	***	0.024	0.131	0.225	< 0.001						
	KLF	0.871	***	0.131	0.615	1.127	< 0.001						
principal components	intercept	- 6.721	***	0.929	-8.540	-4.890	< 0.001	0.575	662.11	1381	975.55	662	1.474
	ln(length)	0.678	***	0.082	0.516	0.838	< 0.001						
	ln(AADT)	0.550	***	0.097	0.358	0.741	< 0.001						
	TramStops	0.188	***	0.023	0.141	0.234	< 0.001						
	C1 (pop.)	0.319	***	0.057	0.205	0.432	< 0.001						
	C2 (eco.)	0.222	***	0.049	0.125	0.317	< 0.001						
land-use + socioeconomic	intercept	-6.832	***	0.946	-8.687	-4.977	< 0.001	0.705	661.86	1396	1018.14	662	1.538
	ln(length)	0.646	***	0.084	0.482	0.810	< 0.001						
	ln(AADT)	0.514	***	0.101	0.317	0.712	< 0.001						
	TramStops	0.194	***	0.025	0.144	0.244	< 0.001						
	VHG 3	0.007	***	0.003	0.002	0.013	< 0.001						
	WoSer	0.044	***	0.012	0.020	0.068	< 0.01						
combined	intercept	-8.044	***	0.898	-9.804	-6.285	< 0.001	0.483	662.23	1367	910.87	662	1.376
	ln(length)	0.705	***	0.081	0.546	0.865	< 0.001						
	ln(AADT)	0.608	***	0.094	0.425	0.791	< 0.001						
	TramStops	0.168	***	0.023	0.123	0.213	< 0.001						
	KLF	0.872	***	0.126	0.625	1.119	< 0.001						
	Sale	0.070	***	0.019	0.033	0.107	< 0.001						

[a] estimation based on adjusted negative binomial distribution
Wald-test of coefficients * p < 0.05; ** p < 0.01; *** p < 0.001

[b] significance of model effects (based on Likelihood-ratio-test)

Beside the exposure (AADT and length) the basic model only includes the number of trams stopping per day on the segment (divided by 100). The other parameters listed in Table 1 either do not prove significant or are highly correlated with environmental parameters associated with greater model effects. For example the dichotomous variable representing parked cars in the segment lost their significant model effects

as soon as any of the three types of environmental variables were included. The number of lanes also does not prove to be significant. Here as well it has to be kept in mind that this parameter correlates highly with AADT.

A comparison of the three different types of environmental variables shows the highest model effects and the best goodness-of-fit when using the adjacent building structure (KLF). The use of the principal components derived from the land-use and socio-economic structure raised in concentric buffers leads to similar results concerning goodness-of-fit, while the initial socio-economic and land-use variables do not reach an effect at this extent. Nevertheless they prove highly significant.

These results suggest that pedestrian traffic can be characterized either with the help of the structure of the built-up area facing the road or the land-use and its occupation within a certain range of influence. However there are differences between the two models that have to be considered. While the coefficients of the length as well as stopping trams have comparable values in both approaches, the AADT coefficient does differ at a noticeable amount (0.668 versus 0.550).

Finally the best goodness-of-fit is achieved by joining both adjacent buildings as well as land-use data in a “combined” model. The selection of the land-use variables is based on the correlation matrix (Table 3) in order to avoid multicollinearity. There are only low correlations between KLF and sales area (Sale), service-related workplaces (WoSer), production-related workplaces (WoPro), as well as school places (Scho). The two latter variables had already proved to have insignificant effects during determination of the land-use model and were therefore discarded. The two remaining variables show similar (highly significant) model effects, so the variable sales area (Sale) is included due to its insignificant correlations with KLF.

The difference in AIC between the “combined” model and the “adjacent building” model of (Δ = 8) can be interpreted as the “adjacent building” model having considerably less empirical support than the “combined” [10]. The “combined” model can be denoted as follows:

$$\mu = 0.00032 \times \text{length}^{0.705} \times \text{AADT}^{0.608} \times e^{(\text{TramStops} \times 0.168 + \text{KLF} \times 0.872 + \text{Sale} \times 0.07)}$$

where:
μ = expected number of accidents (per segment and year)
length = segment length excluding 50 m before and after major intersections (km)
AADT = average annual daily traffic (veh/24h)
TramStops = daily number of trams stopping at stops (100 Trams/24h)
KLF = building structure along the road segment (-)
Sale = density of sales area within 300 m of the segment (m^2/km)

In all five models the coefficient of the exposure length falls below a value of one. Since the segment length is included in the form of a logarithm in the exponentially transformed predictor, a coefficient between zero and one marks a declining increase in accident number with growing length. Subsequently long segments show proportionally fewer accidents than short ones. This effect is likely to have its origin in urban network structures. Segments near the centre of cities tend to be interrupted by intersections and change their characteristics more often than road segments towards the periphery. This happens especially in radial networks. It is likely that higher crash frequencies on short segments therefore depend on the functional complexity of urban central areas. This is assumingly why the coefficients in the environmental models exceed the value estimated in the basic model, as they partly account for the complexity of adjacent land-use. Accordingly, the combined model features the highest coefficient value.

8. DISCUSSION

According to the model results environmental data (socio-economic, land-use, adjacent building structure) have proven to be partly able to substitute missing pedestrian count data within the scope of accident prediction modelling. An exact evaluation of the extent of substitution cannot be carried out since it would require area-wide pedestrian counts.

The two different kinds of data both serve in explaining variance in pedestrian accident counts. While the building structure is easily computed, the environmental data within a range of 300 m affords further statistical treatment in order to account for the correlation structure.

The advantage of using socioeconomic and land-use data lays in its availability insofar as this information is held for traffic demand models. Spread and detail of digital geographical data, moreover, are constantly increasing and a development of more exact procedures for the assignment of these attributes to road networks is likely to produce better estimates of risk factors. Using these spatial data might be a possibility for developing integrated traffic planning tools combining traffic demand modelling and safety assessment at an early stage of planning.

As stated before, principal component analysis assumes the analysed sample as being the entire population. Hence the results can only be extrapolated beyond a specific survey by revealing the same or at least comparable components in further samples. A similar problem arises with the use of statistical modelling in general. The models described are gained by an analysis of one city and therefore depend on its specific structure of land-use. In order to verify these specific results further surveys need to include data from different towns.

REFERENCES

1. Forschungsgesellschaft für Straßen- und Verkehrswesen - FGSV (2006): Hinweise zur Schätzung des Verkehrsaufkommens von Gebietstypen. Cologne: Publications of the Forschungsgesellschaft für Straßen- und Verkehrswesen
2. An, M.; Chen, M. (2007): Estimating Nonmotorized Travel Demand. In: Transportation Research Record, No. 2002, Transportation Research Board, National Research Council, Washington, D. C., pp 18-25
3. Alrutz, D.; Bohle, W. (1999): Flächenansprüche von Fußgängern. Bergisch-Gladbach: Berichte der Bundesanstalt für Straßenwesen; Reihe Verkehrstechnik; Report V 71
4. Monse, A. (2008): Untersuchung der Zusammenhänge zwischen Randnutzung, Sicherheit und nicht motorisiertem Verkehr im städtischen Hauptverkehrsstraßennetz. Dresden: Student research report, Chair of Road Traffic engineering, Technical University Dresden
5. Schueller, H. (2010): Modelle zur Beschreibung des Geschwindigkeitsverhaltens auf Stadtstraßen und dessen Auswirkungen auf die Verkehrssicherheit auf Grundlage der Straßengestaltung. Dissertation, Chair of Road Traffic Engineering, Technical University Dresden. URL: http://www.qucosa.de/fileadmin/data/qucosa/documents/6149/Geschwindigkeitsverhalten_in_Stadtstrassen.pdf
6. Jolliffe, I. T. (2002): Principle Component Analysis (2nd edition). New York: Springer
7. Hutcheson, G.; Sofroniou, N. (1999): The Multivariate Social Scientist – Introductory Statistics Using Generalized Linear Models; London: Sage Publications
8. McCullagh, P.; Nelder, J.A. (1989): Generalized Linear Models (2nd Edition). London: Chapman & Hall
9. Maher, M. J.; Summersgill, I. (1995): A comprehensive methodology for the fitting of predictive accident models. In: Accident Analysis & Prevention, Vol. 28, No. 3, pp 281-296
10. Burnham, K. P.; Anderson, D. R. (2002): Model Selection and Multimodel Inference (2nd edition). New York: Springer